Annotating omission in statement pairs

Héctor Martínez Alonso[1] **Amaury Delamaire**[2,3] **Benoît Sagot**[1]
1. Inria (ALMAnaCH), 2 rue Simone Iff, 75012 Paris, France
2. École des Mines de Saint-Étienne, 158 cours Fauriel, 42000 Saint-Étienne, France
3. Storyzy (Trooclick), 130 rue de Lourmel, 75015 Paris, France
{hector.martinez-alonso,benoit.sagot}@inria.fr
amaury.delamaire@trooclick.com

Abstract

In this piece of industrial application, we focus on the identification of omission in statement pairs for an online news platform. We compare three annotation schemes, namely two crowdsourcing schemes and an expert annotation. The simplest of the two crowdsourcing approaches yields a better annotation quality than the more complex one. We use a dedicated classifier to assess whether the annotators' behaviour can be explained by straightforward linguistic features. However, for our task, we argue that expert and not crowdsourcing-based annotation is the best compromise between cost and quality.

1 Introduction

In a user survey, the news aggregator Storyzy[1] found out that the two main obstacles for user satisfaction when accessing their site's content were redundancy of news items, and missing information respectively. Indeed, in the journalistic genre that is characteristic of online news, editors make frequent use of citations as prominent information; yet these citations are not always in full. The reasons for leaving information out are often motivated by the political leaning of the news platform.

Existing approaches to the detection of political bias rely on bag-of-words models (Zhitomirsky-Geffet et al., 2016) that examine the words present in the writings. Our goal is to go beyond such approaches, which focus on what is said, by instead focusing on what is omitted. Thus, this method requires a pair of statements; an original one, and a shortened version with some deleted words or spans. The task is then to determine whether the information left out in the second statement conveys substantial additional information. If so, the pair presents an omission; cf. Table 1.

Omission detection in sentence pairs constitutes a new task, which is different from the recognition of textual entailment—cf. (Dagan et al., 2006)—because in our case we are certain that the longer text entails the short one. What we want to estimate is whether the information not present in the shorter statement is relevant. To tackle this question, we used a supervised classification framework, for which we require a dataset of manually annotated sentence pairs.

We conducted an annotation task on a sample of the corpus used by the news platform (Section 3). In this corpus, reference statements extracted from news articles are used as long 'reference' statements, whereas their short 'target' counterparts were selected by string and date matching.

We followed by examining which features help identify cases of omission (Section 4). In addition to straightforward measures of word overlap (the Dice coefficient), we also determined that there is a good deal of lexical information that determines whether there is an omission. This work is, to the best of our knowledge, the first empirical study on omission identification in statement pairs.[2]

2 Related work

To the best of our knowledge, no work has been published about omission detection as such. However, our work is related to a variety of questions of interest that resort both to linguistics and NLP.

Segment deletion is one of the most immediate forms of paraphrase, cf. Vila et al. (2014) for a survey. Another phenomenon that also presents the notion of segment deletion, although in a very

[1] http://storyzy.com

[2] We make all data and annotations are freely available at github.com/hectormartinez/verdidata .

different setting, is ellipsis. In the case of an ellipsis, the deleted segment can be reconstructed given a discourse antecedent in the same document, be it observed or idealized (Asher et al., 2001; Merchant, 2016). In the case of omission, a reference and a target version of a statement are involved, the deleted segment in one version having an antecedent in the other version of the statement, in another document, as a result of editorial choices.

Our task is similar to the problem of omission detection in translations, but the bilingual setting allows for word-alignment-based approaches (Melamed, 1996; Russell, 1999), which we cannot use in our setup. Omission detection is also related to hedge detection, which can be achieved using specific lexical triggers such as vagueness markers (Szarvas et al., 2012; Vincze, 2013).

3 Annotation Task

The goal of the annotation task is to provide each reference–target pair with a label: *Omission*, if the target statement leaves out substantial information, or *Same* if there is no information loss.

Corpus We obtained our examples from a corpus of English web newswire. The corpus is made up of aligned reference-target statement pairs; cf. Table 1 for examples. These statements were aligned automatically by means of word overlap metrics, as well as a series of heuristics such as comparing the alleged speaker and date of the statement given the article content, and a series of text normalization steps. We selected 500 pairs for annotation. Instead of selecting 500 random pairs, we selected a contiguous section from a random starting point. We did so in order to obtain a more natural proportion of reference-to-target statements, given that reference statements can be associated with more than one target.[3]

Annotation setup

Our first manual annotation strategy relies on the AMT crowdsourcing platform. We refer to AMT annotators as *turkers*. For each statement pair, we presented the turkers with a display like the one in Figure 1.

We used two different annotation schemes, namely OM_p, where the option to mark an omission is "Text B leaves out some *substantial* information", and OM_e, where it is "Text B leaves out something *substantial*, such as **time**, **place**, **cause**, **people** involved or important **event** information."

The OM_p scheme aims to represent a naive user intuition of the relevance of a difference between statements, akin to the intuition of the users mentioned in Section 1, whereas OM_e aims at capturing our intuition that relevant omissions relate to missing key news elements describable in terms of the 5-W questions (Parton et al., 2009; Das et al., 2012). We ran AMT task twice, once for each scheme. For each scheme, we assigned 5 turkers per instance, and we required that the annotators be Categorization Masters according to the AMT scoring. We paid 0.05$ per instance.

Moreover, in order to choose between OM_p and OM_e, two experts (two of the authors of this article) annotated the same 100 examples from the corpus, yielding the OE annotation set.

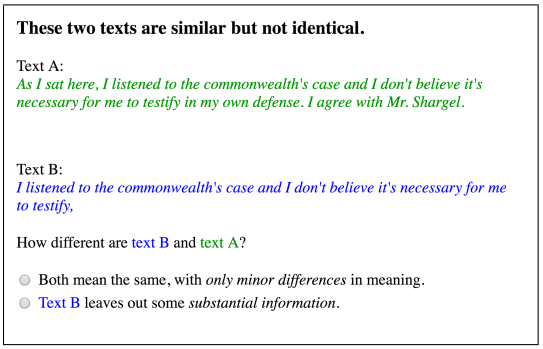

Figure 1: Annotation scheme for OM_p

Annotation results The first column in Table 2 shows the agreement of the annotation tasks in terms of Krippendorff's α coefficient. A score of e.g. 0.52 is not a very high value, but is well within what can be expected on crowdsourced semantic annotations. Note, however, the chance correction that the calculation of α applies to a skewed binary distribution is very aggressive (Passonneau and Carpenter, 2014). The conservativeness of the chance-corrected coefficient can be assessed if we compare the raw agreement between experts (0.86) with the α of 0.67. OM_e causes agreement to descend slightly, and damages the agreement of Same, while Omission remains largely constant. Moreover, disagreement is not evenly distributed across annotated instances, i.e. some instances show perfect agreement, while other instances have maximal disagreement.

We also measured the median annotation time per instance for all three methods; OM_e is almost twice as slow as OM_p (42s vs. 22s), while

[3]The full distribution of the corpus documentation shall provide more details on the extraction process.

Instance	OM_p	OM_e	OE
Example 1 *Interior Minister Chaudhry Nisar Ali Khan on Friday said* no Pakistani can remain silent over the atrocities being committed against the people of the occupied Kashmir by the Indian forces.	0	1	1
Example 2 *I don't feel guilty. I cannot tell you how humiliated I feel.* "I feel robbed emotionally. But we're coming from the east (eastern Europe), we're too close to Russia .."	.8	.2	0
Example 3 The tusks resemble the prehistoric sabre-tooth tiger, but of course, they are not related. *It could make wildlife watching in Sabah more interesting. The rare elephant's reversed tusks might create some problems when it comes to jostling with other elephants. The tusks resemble the prehistoric sabre-tooth tiger, but of course, they are not related*	.6	.4	.5

Table 1: Examples of annotated instances. The 'Instance' column contains the full reference statement, with the elements not present in the target statement marked in italics. The last three columns display the proportion of *Omission* labels provided by the three annotation setups.

Dataset	α	$\tilde{t}$	% Om.	Vote	MACE
Full OM_p	0.52	22	61.72	.65	.63
Full OM_e	0.49	41	63.48	.69	.61
100 OM_p	0.52	22	62.42	.64	.62
100 OM_e	0.54	42	60.00	.61	.58
100 OE	0.67	16	70.87	—	.62

Table 2: Dataset, Krippendorff's α, median annotation time, raw proportion of *Omision*, and label distribution using voting and MACE.

the the expert annotation time in OE is 16s. The large time difference between OM_p and OM_e indicates that changing the annotation guidelines has indeed an effect in annotation behavior, and that the agreement variation is not purely a result of the expectable annotation noise in crowdsourcing.

The fourth and fifth columns in Table 2 show the label distribution after adjudication. While the distribution of *Omission-Same* labels is very similar after applying simple majority voting, we observe that the distribution of the agreement does change. In OM_p, approx. 80% of the *Same*-label instances are assigned with a high agreement (at least four out of five votes), whereas only a third of the *Same* instances in OM_e have such high agreement. Both experts have a similar perception of omission, albeit with a different threshold: in the 14 where they disagree, one of the annotators shows a systematic preference for the *Omission* label.

We also use MACE to evaluate the stability of the annotations. Using an unsupervised expectation-maximization model, MACE assigns confidence to annotators, which are used to estimate the resulting annotations (Hovy et al., 2013). While we do not use the label assignments from MACE for the classification experiments in Section 4, we use them to measure how much the proportion of omission changes with regards to simple majority voting. The more complex OM_e scheme has, parallel to lower agreement, a much higher fluctuation—both in relative and absolute terms—with regards to OM_p, which also indicates this the former scheme provides annotations that are more subject to individual variation. While this difference is arguably of a result of genuine linguistic reflection, it also indicates that the data obtained by this method is less reliable as such.

To sum up, while the label distribution is similar across schemes, the *Same* class drops in overall agreement, but the *Omission* class does not.

In spite of the variation suggested by their α coefficient, the two AMT annotated datasets are very similar. They are 85% identical after label assignment by majority voting. However, the cosine similarity between the example-wise proportions of omission labels is 0.92. This difference is a consequence of the uncertainty in low-agreement examples. The similarity with OE is 0.89 for OM_p and 0.86 for OM_e; OM_p is more similar to the expert judgment. This might be related to the fact that the OM_e instructions prime turkers to favor named entities, leading them to pay less attention to other types of substantial information such as modality markers. We shall come back to the more general role of lexical clues in Section 4.

Given that it is more internally consistent and it matches better with OE, we use the OM_p dataset for the rest of the work described in this article.

4 Classification experiments

Once the manually annotated corpus is built, we can assess the learnability of the *Omission–Same*

decision problem, which constitutes a binary classification task. We aimed at measuring whether the annotators' behavior can be explained by simple proxy linguistic properties like word overlap or length of the statements and/or lexical properties.

Features: For a reference statement r, a target statement t and a set M of the words that only appear in r, we generate the following feature sets:

1. **Dice** (F_a): Dice coefficient between r and t.
2. **Length** (F_b): The length of r, the length of t, and their difference.
3. **BoW** (F_c): A bag of words (BoW) of M.
4. **DWR** (F_d): A dense word representation is word-vector representation of M built from the average word vector for all words in M. We use the representations from GloVe (Pennington et al., 2014).
5. **Stop proportion** (F_e): The proportion of stop words and punctuation in M.
6. **Entities** (F_f): The number of entities in M predicted by the 4-class Stanford Named Entity Recognizer (Finkel et al., 2005).

Table 3 shows the classification results. We use all exhaustive combinations of these feature sets to train a discriminative classifier, namely a logistic regression classifier, to obtain a best feature combination. We consider a feature combination to be the best when it outperforms the others in both accuracy and F1 for the *Omission* label. We compare all systems against the most frequent label (MFL) baseline. We evaluate each feature twice, namely using five-cold cross validation (CV-5 OM_p), and in a split scenario where we test on the 100 examples of OE after training with the remaining 400 examples from OM_p (Test OE). The three best systems (i.e. non-significantly different from each other when tested on OM_p) are shown in the lower section of the table. We test for significance using Student's two-tailed test and $p < 0.05$.

As expected, the overlap (F_a) and length metrics (F_b) make the most competitive standalone features. However, we want to measure how much of the labeling of omission is determined by *which* words are left out, and not just by *how many*.

The system trained on BoW outperforms the system on DWR. However, BoW features contain a proxy for statement length, i.e. if n words are different between ref and target, then n features will fire, and thus approximate the size of M. A distributional semantic model such as GloVe is however made up of non-sparse, real-valued vectors, and does not contain such a proxy for word density. If we examine the contribution of using F_d as a feature model, we see that, while it falls short of its BoW counterpart, it beats the baseline by a margin of 5-10 points. In other words, regardless of the size of M, there is lexical information that explains the choices of considering an omission.

	CV-5 OM_p		Test OE	
	acc.	F1	acc.	F1
MFL	.69	.81	.73	.84
F_a	.79	.81	.76	.83
F_b	.80	.85	.74	.82
F_c	.76	.83	.76	.82
F_d	.74	.84	.76	.84
F_e	.69	.81	.73	.84
F_f	.69	.81	.73	.84
F_{abe}	.83	.87	.74	.81
F_{bf}	.83	.85	.79	.85
F_{cdf}	.81	.86	.82	.88

Table 3: Accuracy and F1 for the *Omission* label for all feature groups, plus for the best feature combination in both evaluation methods. Systems significantly under baseline are marked in grey.

5 Conclusion

We have presented an application-oriented effort to detect omissions between statement pairs. We have assessed two different AMT annotation schemes, and also compared them with expert annotations. The extended crowdsourcing scheme is defined closer to the expert intuition, but has lower agreement, and we use the plain scheme instead. Moreover, if we examine the time need for annotation, our conclusion is that there it is in fact detrimental to use crowdsourcing for this annotation task with respect to expert annotation. Chiefly, we also show that simple linguistic clues allow a classifier to reach satisfying classification results (0.86–0.88 F1), which are better than when solely relying on the straightforward features of different length and word overlap.

Further work includes analyzing whether the changes in the omission examples contain also changes of uncertainty class (Szarvas et al., 2012) or bias type (Recasens et al., 2013), as well as expanding the notion of omission to the detection of the loss of detail in paraphrases. Moreover, we want to explore how to identify the most omission-prone news types, in a style similar to the characterization of unreliable users in Wei et al. (2013).

References

Nicholas Asher, Daniel Hardt, and Joan Busquets. 2001. Discourse parallelism, ellipsis, and ambiguity. *Journal of Semantics*, 18(1):1–25.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In *Machine learning challenges. evaluating predictive uncertainty, visual object classification, and recognising tectual entailment*, pages 177–190. Springer.

Amitava Das, Sivaji Bandyaopadhyay, and Björn Gambäck. 2012. The 5w structure for sentiment summarization-visualization-tracking. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 540–555. Springer.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 363–370. Association for Computational Linguistics.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130.

I. Dan Melamed. 1996. Automatic detection of omissions in translations. In *Proceedings of the 16th Conference on Computational Linguistics - Volume 2*, COLING '96, pages 764–769, Copenhagen, Denmark.

Jason Merchant. 2016. Ellipsis: A survey of analytical approaches. `http://home.uchicago.edu/merchant/pubs/ellipsis.revised.pdf`. Manuscript for Jeroen van Craenenbroeck and Tanja Temmerman (eds.), Handbook of ellipsis, Oxford University Press: Oxford, United Kingdom.

Kristen Parton, Kathleen R McKeown, Bob Coyne, Mona T Diab, Ralph Grishman, Dilek Hakkani-Tür, Mary Harper, Heng Ji, Wei Yun Ma, Adam Meyers, et al. 2009. Who, what, when, where, why?: comparing multiple approaches to the cross-lingual 5W task. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 423–431.

Rebecca J Passonneau and Bob Carpenter. 2014. The benefits of a model of annotation. *Transactions of the Association for Computational Linguistics*, 2:311–326.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar.

Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51th Annual Meeting of the ACL*, pages 1650–1659, Sofia, Bulgaria.

Graham Russell. 1999. Errors of Omission in Translation. In *Proceedings of the 8th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI 99)*, pages 128–138, University College, Chester, England.

György Szarvas, Veronika Vincze, Richárd Farkas, György Móra, and Iryna Gurevych. 2012. Cross-genre and cross-domain detection of semantic uncertainty. *Computational Linguistics*, 38(2):335–367.

Marta Vila, M Antònia Martí, Horacio Rodríguez, et al. 2014. Is this a paraphrase? what kind? paraphrase boundaries and typology. volume 4, page 205. Scientific Research Publishing.

Veronika Vincze. 2013. Weasels, Hedges and Peacocks: Discourse-level Uncertainty in Wikipedia Articles. In *Proceedings of the International Joint Conference on Natural Language Processing*, pages 383–391, Nagoya, Japan.

Zhongyu Wei, Junwen Chen, Wei Gao, Binyang Li, Lanjun Zhou, Yulan He, and Kam-Fai Wong. 2013. An empirical study on uncertainty identification in social media context. In *Proceedings of the 51th Annual Meeting of the ACL*, pages 58–62, Sofia, Bulgaria.

Maayan Zhitomirsky-Geffet, Esther David, Moshe Koppel, Hodaya Uzan, and GE Gorman. 2016. Utilizing overtly political texts for fully automatic evaluation of political leaning of online news websites. *Online Information Review*, 40(3).